Information and the Skewness of Music Sales*

Ken Hendricks
University of Texas at Austin

Alan Sorensen
Stanford University & NBER

September 2008

Abstract

consumers may not have known about them. Our empirical strategy for addressing this issue is based on the effects of new album releases on sales of previous albums by the same artist. The promotional activity and radio airplay associated with a newly released album enhances consumer awareness about the artist, and cause some consumers to discover and purchase the artist's past albums (which are referred to in the industry as "catalog" albums). We call this effect the *backward spillover*. In order to measure it, we constructed a dataset consisting of weekly sales histories for a sample of 355 artists in the period 1993-2002. We observe sales separately for each of the artists' albums, and each artist in the sample released at least two albums (including a debut) during the sample period.

Figure 1 shows two clear examples of the backward spillover. The figure plots the logarithm of weekly national sales for the first and second albums of two popular recording artists, from the time of the artist's debut until six months after the artist's third release. The vertical lines in each graph indicate the release dates of the second and third albums. In the weeks surrounding the release dates, sales of catalog titles increased substantially. In the case of the "Bloodhound Gang," a relatively obscure alternative rock band, the second album was considerably more popular than the first, and its release catapulted sales of the prior album to levels even higher than it had attained at the time of its own release, with the effect persisting for at least three years. For the "Foo Fighters," a more popular hard rock band with a very successful debut album, the impact of the second release was somewhat less dramatic, but still generated an increase in sales of the band's first album. In both examples, the backward spillover is significantly positive for both the second and third album releases.

The first part of our empirical analysis examines the variation in spillover sales in the weeks before and after the new album is released. We use an approach taken from the literature on treatment effects to measure the spillovers. The results confirm that the three patterns illustrated in the above figures hold on average for artists in our sample. First, the increased sales of catalog albums start to appear roughly four weeks *prior* to the release of a new album and increases throughout the pre-release period. Second, the effect peaks in the week of the release and thereafter remains roughly constant as a percentage of sales for many months. Third, the spillovers are larger when the new release is a hit, and especially large when the new release is a hit and the catalog album was not. Finally, we also show that backward spillovers are smaller in an artist's home market (i.e., the city where the artist began her career) even though sales are on average higher in the home market. These patterns suggest that spillovers result from changes in consumers' information. While our

analysis does not rule out explanations based on changes in consumers' utility,[1] the patterns are most easily explained by consumers discovering artists from their new releases and learning about their catalog albums.

We pursue a more structural analysis of the album discovery explanation in the second part of our empirical analysis. We develop and estimate a model of market demand for catalog albums in

boost sales of the existing product (the backward spillover).[4] In a sequel to this paper, Hendricks, Sorensen, and Wiseman (2008) use a variant of the herding models of Banerjee [4], Bikhchandani, Hirshleifer, and Welch [8], and Smith and Sørensen [25] to develop a framework for studying demand for search goods like music albums. Heterogeneous consumers can learn about their preferences for products from the purchasing decisions of other consumers *and* from costly search. The option to search prior to purchasing leads to different market dynamics and outcomes than the standard herding models, and yields testable predictions that are largely consistent with the results of this paper.

More broadly, our paper contributes to a growing literature about the impact of information provision on market outcomes. In markets with a large number of products whose quality is difficult to determine *ex ante*, a variety of mechanisms arise endogenously to provide information to consumers. These mechanisms are typically imperfect, however, and evaluating their impact on what gets sold (and, by extension, what ultimately gets produced) is an important objective for empirical research. Recent papers that address this general topic include Jin and Leslie [20], which examines the effects of publicly posting restaurants' health inspection scores; Sorensen [26], which analyzes the impact of published bestseller lists on the market for books; and Jin, Kato, and List [19], which studies the informational role of professional certifiers in the market for sportscards.

The paper is organized as follows. Section 2 describes the data and provides summary statistics. In Section 3 we use the data to measure the backward spillovers, and document several stylized facts about the spillover. In Section 4 we develop and estimate an album discovery model, and describe the two counterfactual exercises aimed at revealing the quantitative impacts of consumer learning. In Section 5 we discuss alternative explanations. Section 6 concludes.

2 Data

Our data describe the album sales histories of 355 music artists who were active between 1993 and 2002. Weekly sales data for each artist's albums were obtained from Nielsen SoundScan, a market research firm that tracks music sales at the point of sale, essentially by monitoring the cash registers at over 14,000 retail outlets. SoundScan is the principal source of sales data for the industry, and is the basis for the ubiquitous Billboard charts that track artist popularity. Various online databases

[4] In Cabral's paper, for example, the "feedback reputation effect" is exactly analogous to what we call the backward spillover.

were also consulted for auxiliary information (e.g., about genres and record labels) and to verify album release dates.

The sample was constructed by first identifying a set of candidate artists who released debut albums between 1993 and 2002, which is the period for which SoundScan data were available. Sampling randomly from the universe of such artists is infeasible, largely because it is difficult to find information on artists who were unsuccessful. Instead, we constructed our sample by looking for new artists appearing on Billboard charts. The majority of artists in our sample appeared on Billboard's "Heatseekers" chart, which lists the sales ranking of the top 25 new or ascendant artists each week.[5] A smaller number of artists were found because they appeared on regional "New Artists" charts, and an even smaller number were identified as new artists whose debut albums went straight to the Top 200 chart. This selection is obviously nonrandom: an artist must have enjoyed at least some small measure of success to be included in the sample. However, although the sample includes some artists whose first appearance on the Heatseekers list was followed by a rise to stardom, we note (and show in detail below) that it also includes many unknown artists whose success was modest and/or fleeting.[6]

Because our primary objective is to study demand responses to newly released albums, we restrict our attention to major studio releases. Singles, recordings of live performances, interviews, holiday albums, and anthologies or greatest hits albums are excluded from the analysis.[7] The resulting sets of albums were compared against online sources of artist discographies to verify that we had sales data for each artist's complete album history; we dropped any artists for whom albums were missing or for whom the sales data were incomplete.[8] Since timing of releases is an important part of our analysis, we also dropped a small number of artists with albums for which we could not reliably ascertain a release date.[9] Finally, we narrowed the sample to artists for whom we observe

[5] Artists on the Heatseekers chart are "new" in the sense that they have never before appeared in the overall top 100 of Billboard's weekly sales chart—i.e., only artists who have never passed that threshold are eligible to be listed as Heatseekers.

[6] The weekly sales of the lowest-ranked artist on the Heatseekers chart is typically around 3,000, which is only a fraction of typical weekly sales for releases by famous artists who have graduated from the Heatseekers category.

[7] Greatest hits albums could certainly affect sales of previous albums—repackaging old music would likely cannibalize sales of earlier albums—but we are primarily interested in the impact of *new* music on sales of old music. Moreover, there are very few artists in our sample that actually released greatest hits albums during the sample period, making it difficult to estimate their impact with any statistical precision.

[8] The most common causes for missing data were that a single SoundScan report was missing (e.g., the one containing the first few weeks of sales for the album) or that we pulled data for the re-release of an album but failed to obtain sales for the original release.

[9] For most albums, the release date listed by SoundScan is clearly correct; however, for some albums the listed date is inconsistent with the sales pattern (e.g., a large amount of sales reported before the listed release date). In the latter case, we consulted alternative sources to verify the release date that appeared to be correct based on the sales numbers.

the first 52 weeks of sales for at least the first two albums; we then include an artist's third album in the analysis if we observe at least the first 52 weeks of sales for that album (i.e., we include third albums if they were released before 2002).

After applying all of these filters, the remaining sample contains 355 artists and 888 albums. The sample covers three broad genres of music: Rock (227 artists), Rap/R&B/Dance (79 artists), and Country/Blues (49 artists). The artists in the sample also cover a broad range of commercial success, from superstars to relative unknowns. Some of the most successful artists in the sample are Alanis Morissette, the Backstreet Boys, and Shania Twain; examples at the other extreme include Jupiter Coyote, The Weakerthans, and Melissa Ferrick.

Table 1 summarizes various important aspects of the data. The first panel shows the distribution of the albums' release dates separately by release number. The median debut date for artists in our sample is May 1996, with some releasing their first albums as early as 1993 and others as late as 2000. There are 178 artists in the sample for whom we observe three releases during the sample period, and 177 for whom we observe only 2 releases. Note that while we always observe at least two releases for each artist (due to the sample selection criteria), if we observe only two we do not know whether the artist's career died after the second release or if the third album was (or will be)

over the course of their careers. Across the artists in our sample, the simple correlation between first-year sales of first and second releases is 0.52. For second and third releases the correlation is 0.77. Most of an artist's popularity appears to derive from artist-specific factors rather than

prIn toervious truncation our isely be

to measure what happens to sales over time *before* any new album is released.[12]

The regression model is as follows:

$$y_{it} = \beta_0 + \alpha_i + \gamma_t + \sum_{m=2}^{12} \delta_m D_{it}^m + \sum_{s=-13}^{25} \theta_s I_{it}^s + \varepsilon_{it}; \quad (1)$$

where α_i is an artist fixed effect, the γ_t's are time dummies, and the D^m's are month-of-year dummies (to control for seasonality).[13] Here I_{it}^s is an indicator equal to one if the release of artist i's new album was s weeks away from period t, so θ_s measures the new album's sales impact in week s of the treatment window. ($t = 0$ corresponds to the first week following the new release.) Intuitively, after accounting for time and artist fixed effects, we compute the difference in the average sales of album 1 between artists in treatment period s and artists who are not treated for each period, and then average these differences across the time periods. The stochastic error, ε_{it}, is assumed to be heteroskedastic across i (some artists' sales are more volatile than others') and autocorrelated within i (random shocks to an artist's sales are persistent over time). The time dummies (γ_t) allow for a flexible decay path of sales, but implicitly we are assuming that the shape of this decay path is the same across albums. Although differences in the level of demand are captured by the album fixed effects, differences in the shapes of albums' sales paths are necessarily part of the error (ε).

Including separate indicators for successive weeks of treatment allows us to check whether the new release's impact diminishes (or even reverses) over time, which is important for determining whether the effects reflect intertemporal demand shifts. We allow for a 39-week treatment window, beginning 13 weeks (3 months) *before* the release of the new album. The pre-release periods are included for two reasons. First, much of the promotional activity surrounding the release of a new album occurs in the weeks leading up to the release, and we want to allow for the possibility that the backward spillover reflects consumers' responses to these pre-release marketing campaigns. In some cases labels release singles from the new album in advance of the album itself, so that pre-release effects could also reflect advance airplay of the album's songs.[14] Second, including

[12]We believe dropping post-treatment observations is the most appropriate approach, but it turns out not to matter very much: our estimates change very little if we include these observations.

[13]The results reported below are essentially unchanged if we control for seasonality with week-of-year dummies instead of month-of-year dummies.

[14]One might wonder whether the relevant event is the release of the single or the release of the album. Although we have data on when singles were released for *sale*, this does not correspond reliably with the timing of the release on the radio. Radio stations are given advance copies of albums to be played on the air, and a given single may be played on the radio long before it is released for sale in stores. Moreover, even when a single has been released in advance of

pre-release dummies serves as a reality check: we consider it rather implausible that a new album could have an impact on prior albums' sales many months in advance of its actual release, so if the estimated effects of the pre-release dummies are statistical zeros for months far enough back, we can interpret this as an indirect validation of our empirical model.

For the regression described above to yield consistent estimates of the treatment effect, the critical assumption is that the treatment indicators in a period are independent of the idiosyncratic sales shocks in that period. In other words, after controlling for time-invariant characteristics such as genre and artist quality that affect the level of sales in each period, we need the treatment (i.e., the release of a new album) to be random across artists. This is a strong but not implausible assumption. We suspect that the main factor determining the time between releases is the creative process, which is arguably exogenous to time-varying factors. Developing new music requires ideas, coordination, and effort, all of which are subject to the vagaries of the artist's moods and incentives. Nevertheless, the specific question for our analysis is whether release times depend on the sales patterns of previous albums in ways that album fixed effects cannot control.

One possibility is that release times are related to the *shape* of the previous album's sales path. For example, albums of artists that spend relatively more effort promoting the current album in live tours and other engagements will tend to have "longer legs" (i.e., slower decline rates) and later release times than albums of artists that spend more time working on the new album. To check this, we estimated Cox proportional hazard models with time-to-release as the dependent variable, and various album and artist characteristics included as covariates. Somewhat surprisingly, the time it takes to release an artist's new album is essentially independent of the success of the prior album (as measured by first six months' sales) and of its decline rate, after conditioning on genre.[15] These results seem to validate our assumption that release times are exogenous—at least with respect to

$$y_{it} = \tilde{}_0 + \tilde{}_i + \tilde{}_t + \sum_{m=2}^{12} \tilde{}_m D_{it}^m + \sum_{s=13}^{25} \tilde{}_s I_{it}^s + \tilde{}_{it}\ ; \qquad (2)$$

where $y_{it} \equiv y_{it} - y_{it\ 1}$. This model estimates the impact of new releases on the percentage rate of *change* (from y08 -1.os+

significant. Since the dependent variable is the logarithm of sales, the coefficients for specification (1) can be interpreted as approximate percentage changes in sales resulting from the new release. The largest spillover is between albums 2 and 1, with estimates ranging between 40-55%. The spillover of album 3 onto album 1 is smaller, with estimates ranging roughly between 20-38%, and the spillover of album 3 onto album 2 is roughly between 15-35%. Figure 4 shows estimates from specification (2) (the first-differenced model). The solid line plots the cumulative impact implied by the estimated weekly coefficients from the first-differenced model (2), and the dashed lines indicate the 95% confidence bands.[17] The implied effects are qualitatively and quantitatively very similar to those obtained in the undifferenced regressions, which we interpret as reassuring evidence that our results are driven by real effects, not by subtle correlations between current sales flows and the timing of new releases.[18]

In each treatment episode, the estimated impact of the new album three months prior to its actual release is statistically indistinguishable from zero. As discussed above, this provides some reassurance about the model's assumptions: three months prior to the treatment, the sales of soon-to-be-treated albums are statistically indistinguishable from control albums (after conditioning on album fixed effects and seasonal effects). In general, small (but statistically significant) increases start showing up 4-8 weeks prior to the new album's release, growing in magnitude until the week of the release ($t = 0$ in the table), at which point there is a substantial spike upward in sales.

The estimated effects are remarkably persistent: especially for the impact of album 2 on album 1, the spillovers do not appear to be transitory. If the spillover represents consumers who would have eventually purchased the catalog title anyway (i.e., even if the new album were never released), then the coefficients would decline and and eventually would become negative. We have tried longer treatment windows. In some cases, the treatment effect does die out eventually but in none of the cases does the treatment effect turn negative. It is important to note, however, that the increasing coefficients in some specifications do not imply ever-increasing sales paths, since the treatment effects in general do not dominate the underlying decay trend in sales. (In order to save space, the table does not list the estimated time dummies, which reveal a steady and almost perfectly

[17] Because calculating the cumulative impact requires summing coefficients in this specification, the error associated with the cumulative effect at time t reflects the errors of all coefficients up to time t. That is, cumulating the estimates means that the errors cumulate too. Consequently, the confidence bands widen over time.

[18] We also checked the robustness of the estimates by splitting the sample in each treatment based on the median treatment time. As expected, the patterns are the same but the estimated effects are smaller for the albums that are treated early and larger for albums treated later. (This pattern makes sense because our model assumes the effects are proportional: albums treated later will tend to have lower sales flows at the time of treatment, so the proportional impact of the new release will tend to be larger than for albums with high sales flows.) The estimates are always strongly significant.

monotonic decline over time.)

3.3 Spillover Variation

Although it is clear from our results that backward spillovers are significant, it is less clear why the spillovers occur. In this subsection we analyze variation in the magnitudes of the spillovers as a

estimate that weekly sales of her first album more than double when the new album is released. The smallest increase occurs when a hit is followed by a non-hit. The same patterns hold when we examine the impact of the third release on the sales of album 2. The spillovers are large when the new album is a hit, but negligible otherwise. The numbers are slightly smaller than those for the previous album. An important lesson from Table 3 is that although on average (across all types of albums) the backward spillovers are of modest economic significance, they are in fact quite large for the artists that matter: those who have hits or have the potential to produce hits.

In addition to splitting our sample to compare national sales across artists, we can also split the sample geographically to compare sales across markets for a given artist. An especially informative comparison is between an artist's home market (i.e., the city where the artist's career began) and other markets. Because new artists tend to have geographically limited concert tours—in many cases performing only in local clubs—artists in their early careers are more popular in their home markets.

We were able to determine the city of origin for 325 of the 339 artists included in the regression analyses summarized in Figures 3 and 4; 268 of these artists originated in the U.S., so we can observe sales in the home market and compare them to sales in other markets across the nation. SoundScan reports album sales separately for 100 Designated Market Areas (DMAs), each one corresponding to a major metropolitan area such as Los Angeles or Boston. We determined each artist's city of origin, and labeled the nearest DMA to be the artist's home market.[21] It is easy to verify that artists are indeed more popular in their home markets: over 80% of debut albums had disproportionately high sales in the artist's home market, meaning that the home market's share of national first-year sales was higher than the typical share for other artists of the same genre. On average, the home market's share of national sales was 8 percentage points larger than would have been predicted based on that market's share of overall sales within the artist's genre.

Are backward spillovers smaller in artists' home markets? Using the market-level data, we estimate a variant of the regression model in (1):

$$y_{imt} = \ _0 + \ _i + \sum_{g=1}^{4} \ _{gm} G_i^g + \ _1 t + \ _2 t^2 + \ H_{im} + \sum_{k=2}^{12} \ _k D_{it}^k + \sum_{s=\ 13}^{26} I_{it}^s(\ _s + \ H_{im}) +$$

4 An Album Discovery Model

The assumption that most of the spillovers are generated by consumers who have not yet discovered the album seems to us to be a good approximation. Albums are search goods: at low cost, consumers can learn their preferences for an album. They typically learn about an album for free by hearing selected songs played on the radio. Upon hearing the songs, most consumers know whether they like it enough to buy the album or dislike it enough to not buy the album.[25] Those who remain undecided can always learn more by listening to the album online or from friends or at listening posts in record stores. Thus, the binary nature of information in our model—you either know your preferences for an album, or you don't—seems like a reasonable simplification. We discuss the implications of relaxing this assumption in the next section.

The backward spillover arises because the release of a new album generates information about the artist and leads some of the uninformed consumers to discover the artist's catalog albums. In

independent random variables across consumers and release periods; the law of large numbers then implies that proportions of consumers in a large population converge to the associated probabilities.

The proportion of informed consumers for album 1 in period t accumulates according to the equation

$$q_{1;t} = q_{1;t-1} + (1 - q_{1;t-1})\frac{ae^{bS_{t;t}}}{(1-a) + ae^{bS_{t;t}}}; \tag{4}$$

where $q_{1;t}$ denotes the proportion of consumers who know their preferences for album 1 at the end of period t. For period 1, we set $q_{1;0} \equiv q_0$, so q_0 is interpreted as the baseline awareness of the artist prior to her debut.

The probability that consumer i purchases album 1 in period t conditional on discovering the album is simply denoted p_1. We make the critical assumption that a consumer's utility for album 1 does not change over the release periods. Because the choice set is changing over the release periods, this assumption requires preferences to be additive across albums. Additivity is a strong assumption, but it is testable, as we explain below. Notice that we are also implicitly assuming that the discovery probability is independent of the consumer's preferences.

Since preferences are assumed not to change across release periods, spillover sales reflect changes in the number of informed consumers: $(q_{1;t} - q_{1;t-1})N$, where N is the number of potential consumers for the album. Appealing to the law of large numbers, sales of album 1 in period $t > 1$ are given by

$$S_{1;t} = p_1(q_{1;t} - q_{1;t-1})N : \tag{5}$$

Sales of album 1 in its own release period are simply $S_{1;1} \equiv p_1 q_{1;1} N$. Since $q_{1;1}$ is a function of $S_{1;1}$ (as indicated in equation (4)), sales are reinforcing: higher sales lead to more consumers discovering the album, which further increases sales.[30] This, along with the fact that p_1 is unob-

period. We make this assumption primarily for convenience. However, we tested this assumption by estimating models in which a was allowed to be different in period 1 vs. period 2, or in which b was allowed to be different in period 1 vs. period 2. In neither case could we reject the hypothesis that the parameter is equal across periods.

[30] It is straightforward to show that a solution (i.e., fixed point) to this sales relationship always exists, that there are either one or three solutions (generically), and the minimum and maximium solutions are increasing in album quality (i.e., p_k). Multiple equilibria can arise because of the logistic learning curve and the lack of coordination among radio stations in choosing playing time. However, for the learning curve we estimate below, it turns out that the fixed point is unique for every album.

servable, makes it difficult to estimate the parameters of the model using only data on sales of the album in its own release period.

Instead, we estimate the parameters of the model using the spillover of album 2 onto album 1. Specifically, our estimation exploits the comparison of album 1 sales in release period 2 to album 1 sales in release period 1. There are two main reasons for this approach. First, sales of the *new* album in release period 2 shift consumer awareness ($q_{1;2}$) exogenously (conditional on sales of album 1 in period 1). Second, the comparison allows us to eliminate unobservable p_1 from the model: substituting $p_1 = S_{1;1}/q_{1;1}N$ into equation (5), we obtain

$$\frac{S_{1;2}}{S_{1;1}} = \frac{q_{1;2} - q_{1;1}}{q_{1;1}} = \frac{(1 - q_{1;1})}{q_{1;1}} \frac{ae^{bS_{2;2}}}{(1 - a) + ae^{bS_{2;2}}} : \tag{6}$$

Notice that N, the (unknown) number of potential consumers for album 1, is also eliminated in this step. The remaining unobservable is $q_{1;1}$, the fraction of consumers who know their preferences for album 1 in release period 1. Substituting for this variable (using equation (4)) and taking logs, the spillover equation becomes

$$\log \frac{S_{1;2}}{S_{1;1}} = \log \left(\frac{(1 - q_0)(1 - a)}{q_0(1 - a) + ae^{bS_{1;1}}} \right) + \log(a) + bS_{2;2} - \log \left(1 - a + ae^{bS_{2;2}}\right) : \tag{7}$$

In examining the backward spillovers, we noticed that their magnitude appears to decrease as a function of time between releases. To accommodate this feature of the data, we make an *ad hoc* modification to equation (7) that allows for depreciation. We assume that mean utility declines over time at a rate that is common across albums, writing the purchase probability for a consumer who learns her preferences for album 1 in period $t > 1$ as

$$p_{1;t} = p_1 e^{\quad T_{1;t}} \tag{8}$$

where $T_{1;t}$ is the length of time between the release of album 1 and the release of album t. This specification allows consumers to have a taste for "newness" and for the spillover to decline as a function of time between releases. Using equation (8), the spillover equation that we take to the data is

$$\log \frac{S_{1;2}}{S_{1;1}} = \log \left(\frac{(1 - q_0)(1 - a)}{q_0(1 - a) + ae^{bS_{1;1}}} \right) - \quad T_{1;2} + \log(a) + bS_{2;2} - \log \left(1 - a + ae^{bS_{2;2}}\right) + \quad ; \tag{9}$$

where is the error term.

It will be convenient to standardize the length of the period over which to measure albums' sales. We calculate sales over a one-year period: $S_{1;1}$ is measured as first-year sales of album 1, and $S_{1;2}$ is measured as cumulative sales of album 1 during the first year of release 2. The definition of a release period as one year is long enough for the sales dynamics to have run their course: almost anyone who was going to learn about the new release and buy it before the release of the next album will have done so within the first year. It introduces some measurement error into the model, since the fraction of informed consumers at the end of an album's first year is not the same as the fraction of informed consumers at the time of the next release. However, the error is small: on average, first year sales represent 85% of cumulative sales at the time of the next release.[31] Time between releases ($T_{1;2}$

4.2 Results

Before reporting our parameter estimates, we explain briefly how they are identified by the data. Our estimate of b is driven by the sensitivity of the backward spillovers to sales of the new album. As shown in Table 3 above, backward spillovers are significantly larger when the new release is successful, so we should expect a positive estimate of b. If instead spillovers were invariant to the success of the new release, then we would expect our estimate of b to be close to zero. We also found that sometimes backward spillovers occur even when the new release sells very little. The observed magnitude of backward spillovers in such cases identifies a, the baseline flow of learning. If a were zero, we would expect backward spillovers *only* when the new release is successful; the higher is a, the larger the spillovers will be even in instances where the new release is a dud. The q_0 parameter is identified by the average magnitudes of backward spillovers. If q_0 is near zero, the model allows for large spillovers that may depend on sales of the new album (through the a and b parameters); if instead q_0 approaches 1, the model predicts very small spillovers no matter how strong are the sales of the new album. Notice that unlike the a parameter, q_0 does not interact with sales of the new album. This is what allows the two parameters to be separately identified, in spite of serving similar purposes in the model. Finally, is identified by the extent to which spillovers tend to be smaller when the time between releases is longer. (This pattern in the data is precisely what motivated the inclusion of the term.)

The first column of Table 5 reports nonlinear least squares estimates of equation (9) based on national sales.[33] Our estimate of q_0 implies that on average 18 percent of potential buyers are

market. We read this comparison as offering basic support for our interpretation of the model's parameters. Of course, it also suggests there may be other interesting sources of heterogeneity in learning across markets. In the discussion that follows, however, we simply focus on the estimates based on national sales, leaving market-level heterogeneity as an issue to explore more fully in future work.

Figure 5 illustrates the discovery function defined by the first column of Table 5. Its shape is determined by the parameters *a* and *b*, with *a* representing the baseline learning rate, and *b* representing the rate at which learning increases with sales. Initially, the probability of discovery increases at an increasing rate as a function of sales; but eventually the function becomes concave and the fraction of informed consumers approaches one. The inflection point is at 2.54 million sales. As noted above, the logistic learning curve can potentially give rise to multiple equilibria; however, this turns out to be irrelevant given our parameter estimates.[34]

Our estimates imply that learning is nearly complete for artists with extremely successful albums. For example, an artist whose debut album sells 10 million copies—which would classify it as a huge hit, and earn it the RIAA "Diamond" award—would be known to 99% of consumers.[35] At the other end of the success spectrum, the majority of consumers remain uninformed: if a debut album sells fewer than 500,000 copies in the first year, our estimates suggest only a third of potential consumers will have discovered the artist in that year. Note that the most successful debut album in our dataset sold roughly 8.2 million copies in its first year, so the graph shown does not extrapolate far beyond the range of the data.

The estimated value of *a* suggests that with each new album released, at least 16 percent of previously uninformed consumers will learn their preferences for the catalog album even if the new album has zero sales. Because learning is cumulative in our model, this implies (somewhat counterintuitively) that an artist could become a household name by releasing a long sequence of very low-quality albums. However, the numbers imply that such an artist would need to release 13 such albums before 90% of consumers would become aware of the first album. By contrast, a successful artist can become famous with only two or three hit albums. For example, after a sequence of three "triple-platinum" albums (sales of 3 million each), 94% of consumers would know their preferences for the first such album.

[34] Due in particular to the relatively high estimated values of q_0 and a, for the albums in our sample the relationship described by $S_{1;t} = p_1 q_{t;t} N$ has only one fixed point.

[35] This need not mean 99% of *all* consumers, but rather 99% of the relevant population of consumers, which presumably consists of those consumers who could potentially be exposed to information about the album. The model implicitly defines the size of the market by the point at which there can be no backward spillovers.

As mentioned above, the model assumes that preferences are additive. In particular, preferences for a given album do not depend on the existence or characteristics of other albums, even by the same artist. We can test this assumption by looking at the spillover sales of album 1 after the release of album 3. If discovery is uncorrelated with preferences and preferences do not change over time, additivity implies that the fraction of new consumers who buy album 1 in period 2 is the same as the fraction who buy album 1 in period 3 (controlling for age effects as indexed by ^). Specifically, the relationship between the two spillovers is given by

$$\log \frac{S_{1;2}}{S_{1;3}} = \hat{}(T_{1;3} - T_{1;2}) + \log\left(\frac{q_{1;2} - q_{1;1}}{q_{1;3} - q_{1;2}}\right): \tag{10}$$

Given parameter estimates ($\hat{q}_0$,$\hat{a}$, $\hat{b}$, ^), we can calculate the fraction of consumers who knew their preferences for album 1 at the ends of periods 1, 2, and 3 (i.e., $\hat{q}_{1;1}$, $\hat{q}_{1;2}$, and $\hat{q}_{1;3}$, respectively), and plug these in to the right-hand side of equation (10) to obtain a prediction for $\log(S_{1;2}=S_{1;3})$. Comparing our predictions to what we observe in the data, we find that the differences are on average positive. (The average difference is .291, with a standard error of .089.) In other words,

which is a function of unobservable preferences. However, we can use the same trick as above to

denote the counterfactual sales of album 2 if it had instead been the debut album, and let $\tilde{q}_{1;1}$ be the fraction of consumers who would have discovered album 2 in that case. Then

$$\tilde{S}_2 = p_2 \tilde{q}_{1;1} N = S_{2;2} \frac{\tilde{q}_{1;1}}{q_{2;2}}; \tag{12}$$

where the second equality follows from the fact that $S_{22} = p_2 q_{2;2} N$. The equation states that we can estimate counterfactual sales of album 2 by simply rescaling the observed sales of album 2 by a factor equal to the ratio of $\tilde{q}_{1;1}$ to $q_{2;2}$.

Since $\tilde{q}_{1;1}$ is itself a function of $\tilde{S}_2$

where the second equality follows from the fact that observed sales $S_{1;1}$ are equal to $p_1 q_{1;1} N$, and we calculate $\hat{q}_{1;1}$ from equation (4) using our estimates of the learning parameters ($\hat{q}_0$, $\hat{a}$, and $\hat{b}$, as reported in Table 5). The idea is very simple: for an album that sold 100,000 units in its release period, if we estimate that $q_{1;1}$ was 33.3% for that album (i.e., only a third of consumers knew about it), then the implied counterfactual sales for that album would be 300,000.

The results of these calculations are summarized in Table 6. Our estimates imply that albums at the very top are *not* substantially undersold. Almost all consumers learn about a major hit (such as Alanis Morissette's *Jagged Little Pill*) in its release period, and the counterfactual is only a small change from reality. Albums at the bottom end of the success spectrum are also not undersold, but for a different reason. Even though most consumers are unaware of these albums, the albums' qualities are sufficiently low that sales would be minimal even if everyone were fully informed. By contrast, we estimate that moderately successful (but sub-superstar) albums are substantially undersold, in the sense that many would-be buyers remain uninformed about such albums. For all but the very top artists, album sales would have doubled or even tripled if every consumer had been aware of the album, and in absolute terms these differences would have been very large for the moderately successful artists.

Of course, one might argue that some of these potential sales would have occurred when later releases by the artist generated new information. However, consumer utility for an album appears to decline over time, so delays in discovery times are costly. Furthermore, substantial learning takes place only if one of the later releases is a major hit. If we look at sales of debut albums in the four years following their release years, we find that the typical artist recovers only 25% of the "lost sales" implied by our counterfactual analysis.[38] For artists who had a major hit (defined as an album selling over 2 million units in its first year) on a subsequent release, typically 45% of the lost sales are recovered. For example, the debut album from Coal Chamber (noted in Table 6) sold fewer than 1 million units in its first year, but over 3 million units after that, largely because the band's second release was a major hit. So the eventual sales of the debut album were even greater than what our model predicted for counterfactual sales. In contrast, Queen Pen's debut album sold over 3 million units in its first year, but less than a quarter million after that, perhaps because her second release was not very successful. Our data indicate that most artists experience a fate similar to Queen Pen's.

A central implication of our finding that lesser-known artists' albums are undersold is that com-

[38] In making this calculation we restricted our attention to the 190 artists for whom we observe at least 5 complete years of sales data for the first album.

mercial success in this industry is more concentrated than it would be in a world where consumers were more fully informed. More popular albums are more widely promoted, so more consumers know about them, and popularity is self-reinforcing. The fact that artists release multiple albums only serves to amplify the skewness in sales: popular first albums are likely to be followed by popular second albums, and the effect of the consequent backward spillovers will tend to increase sales disproportionately for albums that were already popular.[39] Our model does not merely measure this direct effect of spillovers on album sales; instead, it uses the spillover as a way of estimating the extent to which sales are dependent on information.

Exactly how much of the observed skewness in album sales can be attributed to consumers' lack of information about the choice set of albums? Our model cannot address this question directly unless we assume that preferences are additive. This approximation is acceptable when the counterfactual consists of adding an album to the actual choice sets of uninformed consumers, but it is not very plausible when the counterfactual consists of giving all consumers the choice set of *all* albums. Nevertheless, our estimates of counterfactual sales do provide an interesting benchmark. Figure 6 plots the distribution of counterfactual first-year sales ($\tilde{S}_{1;1}$, as defined above) in comparison to the observed sales of debut albums ($S_{1;1}$). The counterfactual distribution of sales is still quite skewed, but it is substantially less concentrated than the distribution of actual sales. The Gini coefficient is .647, as opposed to .724 for actual sales.[40] Among the artists in our sample, fewer than half (48%) sold more than 100,000 units of their debut albums, but our estimates imply that if consumers had been fully informed then nearly three quarters (72%) would have met or exceeded this threshold.

We suspect that allowing for substitution effects would make the flattening of the distribution even more pronounced. To understand why, note that making consumers fully informed about all albums would have two opposing effects on the sales of each album. The direct effect would be an increase in sales resulting from more consumers knowing about the album. The indirect effect would be a decrease in sales due to consumers becoming more aware of competing albums. For the most successful of albums, the direct effect is small (most consumers know about these albums already), so the indirect effect is likely to dominate. For the least successful albums, the increased competition would come mostly from albums of higher quality, so sales would likely decrease for these albums too. For albums in the middle (i.e., moderately successful albums), the direct effect is large (since we estimate that consumer awareness is still somewhat low for such albums), and the

[39] The fact that artists release multiple albums also means that the distribution of success across artists will be even more skewed than the distribution across albums.

[40] As a point of reference, the income distribution in the United States has a Gini coefficient of around .47.

the discovery probability function from any one of the three spillovers (i.e., 2-1, 3-1, or 3-2), and use the results to test our model.

One alternative would be a model in which learning is more gradual and nuanced. Instead of making the simplifying assumption that consumer knowledge is binary, we could instead assume that consumers update their preferences gradually as their hear more songs from an album on the radio. In this model, consumers may know about an album but still need to hear more songs from the artist before deciding whether they'll like it enough to buy it. Spillovers occur when a new release generates information that convinces a significant proportion of these consumers that the album is worth purchasing. If higher sales is "good news" about album quality, then each consumer's probability of purchasing the catalog album will be an increasing function of sales of the new release. We could specify a functional form for this probability—similar to the one that we specified for the probability of discovery—and estimate its parameters using, say, the 2-1 spillover. The main problem in taking this model to the data, however, is that consumers who did not purchase the catalog album in prior release periods are not a random sample: they are less likely to buy the album than a randomly selected, uninformed consumer. While it would be possible in principle to model the selection process, the resulting model would be much more complicated than the one we estimated above. More importantly, the selection issue would make our results highly sensitive to the functional forms chosen for the distributions of signals and tastes.

Other models could attribute the backward spillovers to changes in utility rather than changes in information. If consumers have supermodular preferences over albums by the same artist, for example, then a new release increases the utility of the artist's catalog albums. Consumers who were previously not willing to buy a catalog album may do so when they can consume it together with the new album.[42] Another form of consumption complementarities is social effects: a consumer's utility from an album may depend on the number of other consumers purchasing the artist's albums.[43] A new release that sells well could increase utility (and hence sales) of catalog albums if (a) the social effects operate to some extent at the artist level and (b) the social effects associated with the new album exceed the social effects generated by the catalog albums themselves when they were released. The implications of these models are similar to those of a non-binary learning model: sales of the new release affect the purchasing probabilities of consumers who have not yet

[42]The complementarity could be interpreted as a characterization of fans: e.g., when consumers listen regularly to an artist's music, they become accustomed to it or invested in the image associated with it, and therefore more likely to purchase more music from that artist. Such complementarities would be similar to those modeled by Becker, Grossman, and Murphy [5] to describe cigarette addiction, and by Gentzkow [15] to describe consumption of online and print editions of a newspaper.

[43]See Becker and Murphy [6] and Brock and Durlauf [9] for insightful overviews of models with social effects.

purchased the catalog album, and the latter are a selected sample. In this case, the magnitude of the spillover and its variation across release periods depends on the structure of consumer preferences and how these preferences are distributed in the population.

We cannot definitively reject these alternative models using aggregate sales data. We have tried estimating versions of these models, and have found that they have difficulty rationalizing the variation in spillovers across artists and release periods. For example, the selection effect makes it difficult to obtain significant spillover from later releases of successful artists. Part of the appeal of the binary learning model (aside from its plausibility) is that it provides a unified model of spillovers across artists and album-pairs that, somewhat surprisingly, fits the data very well. This finding suggests to us that while gradual learning, album complementarities, and social effects may play a role, the main factor determining demand for albums is whether consumers know about them and the process through which they obtain this knowledge (i.e., radio play).

Preference-based models also have different implications about how changes in the market environment will affect the distribution of album sales. For example, as explained in the previous section, recent evidence suggests that internet technologies have led to a flattening of the distribution of music sales. This is a natural implication of the album discovery model. By contrast, even though a model based on complementarities between albums could possibly explain the backward spillovers, it would not have direct implications about the impact of internet technologies that facilitate the flow of information. In a model based on social effects, the impact of internet music technologies would be ambiguous; but intuitively one might expect such a model to predict an *increase* in skewness, since the internet makes it much easier for consumers to observe each others' purchases and coordinate on what is popular.

5 Conclusion

We have shown that the release of a new album generates substantial, persistent increases in the sales of previous albums by the same artist. The evidence strongly suggests that these backward spillovers are generated by changes in information: a new album release causes some consumers to discover artists and albums about which they were previously uninformed. Cross-sectional variation in the spillovers allows us to make quantitative inferences about the importance of product discovery and its impact on market outcomes. Estimates of our model imply that the distribution of sales is substantially more skewed than it would be if consumers were more fully informed.

In particular, mid-range artists' albums are dramatically undersold (to the tune of hundreds of

(in the ex post sense) in buying albums than, say, books, and hit albums are less likely to be oversold than bestselling novels.[50] Another example is personal computers: Goeree [16] argues that the rapid pace of technological change in computers leads consumers to be less than fully informed about the set of available products. Our findings indicate that the distribution of success in these markets may be very different from what it would be in a world with more fully informed consumers.

[50]This may partly explain why book sales are much more skewed than music sales. (See Sorensen [26] for some evidence and discussion of the skewed distribution of sales for hardcover fiction.)

References

[1] Ackerberg, D. (2001), "Empirically Distinguishing Informative and Prestige Effects of Advertising," *RAND Journal of Economics*, 32(2).

[2] Ackerberg, D. (2003), "Advertising, Learning, and Consumer Choice in Experience Good Markets: An Empirical Examination," *International Economic Review*, 44(3).

[3] Anderson, C. (2006), *The Long Tail: Why the Future of Business is Selling Less of More*, Hyperion (New York).

[4] Banerjee, A. (1992), "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107(3), pp. 797-817.

[5] Becker, G., Grossman, M., and K. Murphy (1994), "An Empirical Analysis of Cigarette Addiction," *American Economic Review*, 84(3), pp. 396-418.

[6] Becker, G., and Murphy, K. (2000), *Social Economics: Market Behavior in a Social Environment*, Harvard University Press.

[7] Benkard, C. L. (2000), "Learning and Forgetting: The Dynamics of Aircraft Production," *American Economic Review*, 90(4), pp. 1034-1054.

[8] Bikhchandani, S., Hirshleifer, D. and Welch, I. (1992), "A Theory of Fads, Fashions, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100, pp. 992-1026.

[9] Brock, W. and S. Durlauf (2001), "Interactions-Based Models," in *Handbook of Econometrics*, v1ld.932 us932 0 Td (,)-250(t212(JTd (,)-H40(monomic)-.)-273,)-2 [([5l399((dc5l399(eds2(JTd (,Am

[16] Goeree, M. S. (2005), “Advertising in the U.S. Personal Computer Industry,” Claremont McKenna working paper.

Table 1: Summary Statistics

				Percentiles		
	N	Mean	Std. Dev.	.10	.50	.90
Date of release:						
album 1	355	13may1996	102	22aug1993	05may1996	28feb1999
2	355	20jul1998	108	23jul1995	02aug1998	27may2001
3	178	03jun1999	90	13oct1996	04aug1999	05aug2001
First year sales:						
album 1	355	312,074	755,251	7,381	78,360	781,801
2	355	367,103	935,912	10,705	55,675	951,956
3	178	450,716	867,630	7,837	71,674	1,461,214
overall	888	361,864	854,420	9,095	67,558	996,460
First 4 weeks / First year:						
album 1	355	.121	.111	.016	.085	.265
2	355	.263	.137	.086	.263	.441
3	178	.305	.131	.134	.305	.500
overall	888	.214	.148	.031	.198	.419
Peak sales week:						
album 1	355	31.9	47.8	0	15	87
2	355	7.83	23.1	0	0	28
3	178	4.05	13.1	0	0	12
overall	888	16.7	36.3	0	1	46
Weeks between releases:						
1 & 2	355	114	53.5	58	107	179
2 & 3	178	111	46.7	58	104	169

Table 2: Seasonality in release dates

	Percent of releases occurring			
Month	Album 1 (*n*=355)	Album 2 (*n*=355)	Album 3 (*n*=178)	Overall (*n*=888)
Jan	3.94	3.10	3.37	3.49
Feb	8.17	4.23	3.93	5.74
Mar	13.24	9.58	11.80	11.49
Apr	9.01	8.45	8.99	8.78
May	11.83	9.01	7.30	9.80
Jun	7.61	12.68	6.74	9.46
Jul	8.45	9.01	10.11	9.01
Aug	11.55	9.58	10.67	10.59
Sep	7.32	11.27	11.80	9.80
Oct	12.39	10.70	16.29	12.50
Nov	5.92	11.83	6.74	8.45
Dec	0.56	0.56	2.25	0.90

Table 3: Spillovers and hits

Album 1, Album 2:	Hit, Hit	Hit, Not	Not, Hit	Not, Not
N	53	45	34	206
Median # weeks to release 2	108	124	101	104
Median weekly sales (album 1) prior to release:	1,888	318	342	154
Median weekly decline around release:	-0.021	-0.018	-0.018	-0.011
Estimated total change in sales:	22,161	660	14,557	883
Percentage change in sales:	42.7	7.2	148.5	17.6
Average of (sales before next release)/(first 4 years' sales):	0.73	0.85	0.55	0.62
Album 2, Album 3:	Hit, Hit	Hit, Not	Not, Hit	Not, Not
N	49	13	12	99
Median # weeks to release 3	105	117	95	103
Median weekly sales (album 1) prior to release:	1,555	466	844	85
Median weekly decline around release:	-0.013	-0.026	0.004	-0.010
Estimated total change in sales:	19,884	1,110	20,788	687
Percentage change in sales:	40.6	9.5	56.4	24.6
Average of (sales before next release)/(first 4 years' sales):	0.73	0.84	0.59	0.65

Hits are defined as albums that sold over 250,000 units nationally in the first year. Albums that didn't clear this threshold are the "Not" albums (i.e., not hits). The estimated total changes and percentage changes in sales reflect increases over the 39-week treatment window.

Table 4: Sales and spillovers in the artist's home market

	2→1	3→2
Home market (^)	0.814	0.647
	(0.006)	(0.008)
Home market × new release period (^)	-0.105	-0.104
	(0.010)	(0.013)
# observations	2,727,890	1,437,340
# artists	268	142

Estimates of the regression model described in equation (3); the dependent variable is log sales. ^ measures the average difference in log sales between the artist's home market vs. other markets, and ^ measures the average difference in the backward spillover in the artist's home market vs. other markets. Other coefficients are omitted to save space.

Table 5: Estimated parameters of album discovery model

			DMA-level sales	
		National sales	Home market	Non-home markets
"Baseline" awareness:	q_0	0.180	0.305	0.143
		(0.049)	(0.052)	(0.045)
Learning function parameters:	a	0.161	0.284	0.144
		(0.072)	(0.121)	(0.061)
	b	0.065	0.112	0.098
		(0.013)	(0.027)	(0.018)
Time between releases:		0.701	0.567	0.697
		(0.073)	(0.086)	(0.086)
	N	311	247	247
	R^2	.288	.214	.274

Asymptotic standard errors in parentheses.

Table 6: Counterfactual sales under full information

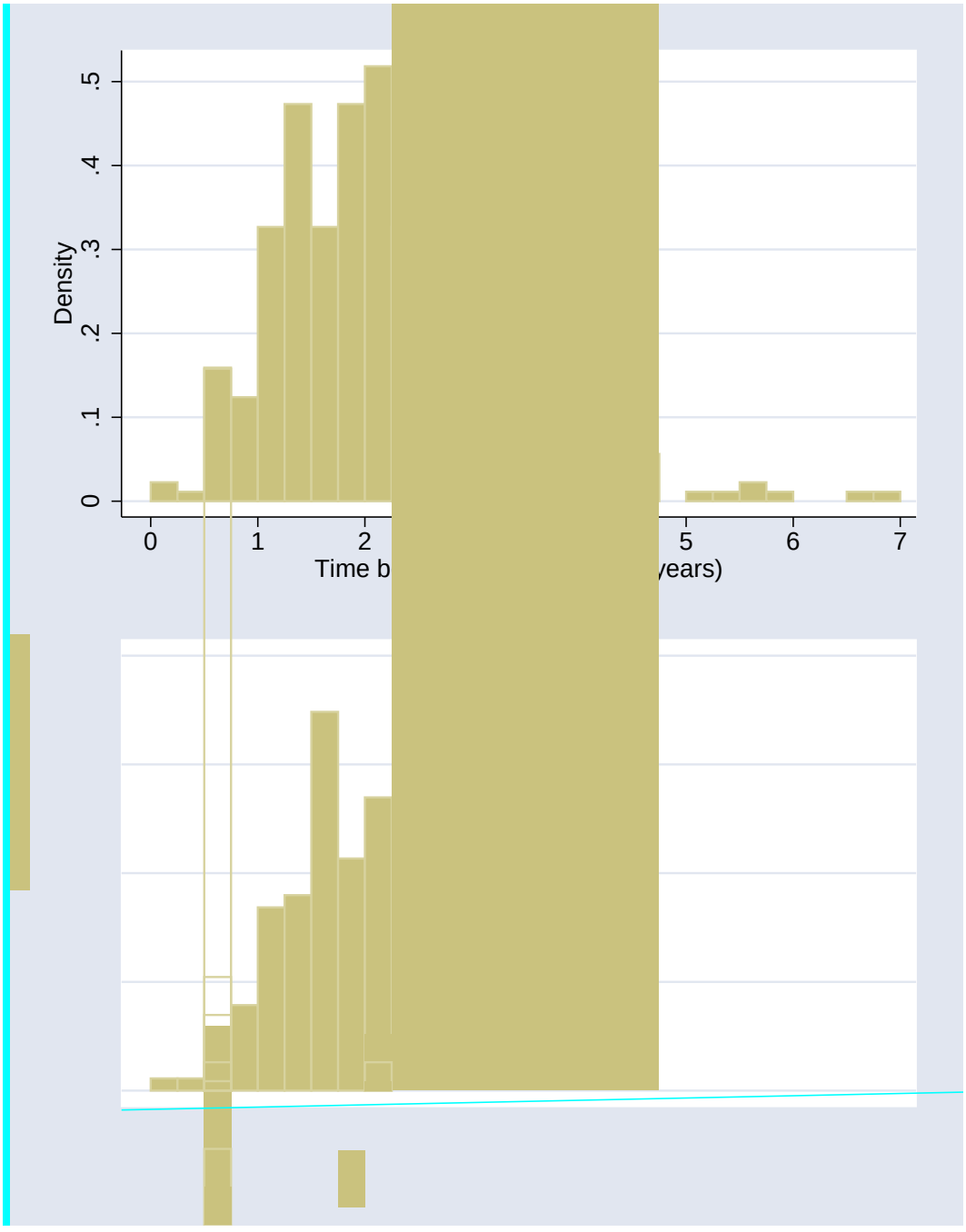
Percentile	Artist/Title	Observed sales	Sales if $q_{1;1}$=1	Difference
max	Alanis Morissette/Jagged Little Pill	8,204,835	8,378,094	173,259
.90	Coolio/It Takes a Thief	802,380	2,113,075	1,310,695
.75	Queen Pen/My Melody	302,254	902,572	600,318
.50	Coal Chamber/Coal Chamber	90,449	284,108	193,659
.25	Wild Colonials/Fruit of Life	27,323	87,095	59,772
.10	Prince Paul/Psychoanalysis...	8,193	26,232	18,039
min	Oleander/Shrinking the Blob	883	2,832	1,949

Compares first-year sales of debut albums to the model's prediction of sales if consumers had been fully informed about the album (i.e, if $q_{1;1}$ had been equal to 1 for that album).

Figure 1: Album sales paths for two examples

These graphs show log(sales) over time (measured in weeks) for the artists' first and second albums. The vertical lines indicate the release dates of albums 2 and 3. The graphs illustrate the *backward spillover*: the release of a new album tends to cause a sales increase for previous albums by the same artist.

Figure 2: Distributions of Elapsed Time Between Releases



The upper panel plots the elapsed time between the releases of albums 1 and 2 by the 355 artists in our sample. The lower panel plots time between releases 2 and 3 for the 178 artists for whom we observe a third album.

Figure 3: Time patterns of backward spillovers

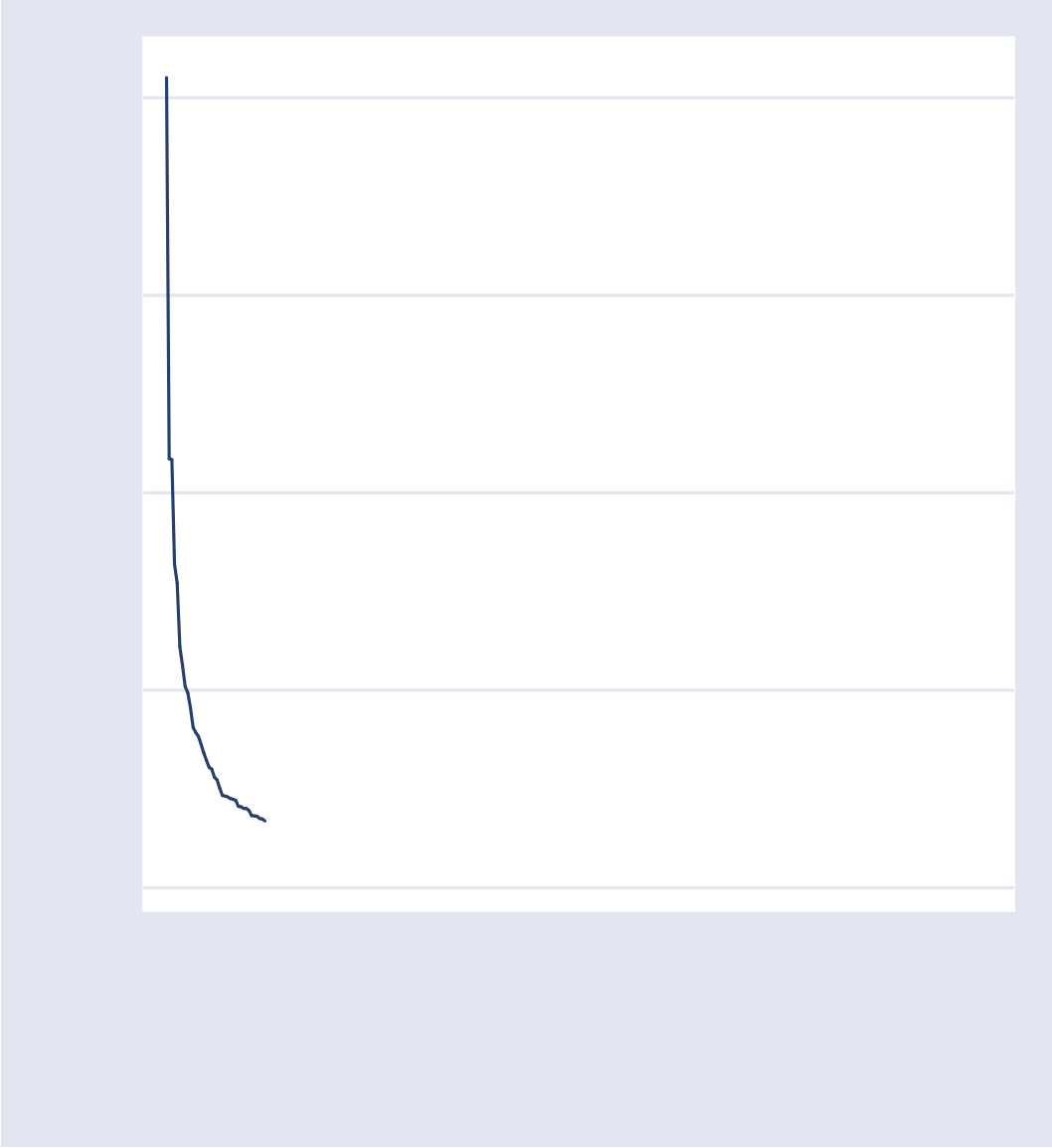
Figure 4: Time patterns of backward spillovers: first-differences model

These graphs are analogous to those in Figure 3, except that the coefficients are estimated from the first-differenced model (equation (2)). The confidence intervals (dotted lines) expand over time because the effect at time *t* is the sum of all coefficients up to time *t*; so the error associated with the cumulative effect at time *t* reflects the errors of all coefficients up to time *t*.

Figure 5: Estimated album discovery function

This graph shows our estimate of equation (4) for the debut album—i.e., the fraction of the relevant population of consumers who discover album 1, as a function of its sales. For an album with sales near zero, we estimate that roughly 30% of consumers will know about it. For an album with sales over 8 million, we estimate that roughly 98% of consumers know about it.

Figure 6: Counterfactual sales distribution for debut albums



The solid line shows the actual first-year sales of the top 300 debut albums in our sample plotted against sales rank. The dotted line indicates our estimate of what this plot would look like if all consumers were fully informed and preferences were additive across albums (i.e., no substitution effects).